LOG Analiz: Erişim Kayıt Dosyaları Analiz Yazılımı ve GOP Üniversitesi Uygulaması

Turgut ÖZSEVEN[1], Muharrem DÜĞENCİ[2]

[1]Turhal Meslek Yüksekokulu, Gaziosmanpaşa Üniversitesi, Tokat, Türkiye
[2]Endüstri Mühendisliği Bölümü, Karabük Üniversitesi, Karabük, Türkiye
turgut@gop.edu.tr, mdugenci@karabuk.edu.tr
(Geliş/Received: 22.12.2010; Kabul/Accepted: 20.03.2011)

Özet— İnternet kullanıcılarının web sitesi ziyareti süresince geride bıraktığı izler sunucu üzerindeki erişim kayıt dosyalarında tutulmaktadır. Bu verilerin analiz edilerek bilgiye dönüştürülmesi web madenciliği ile yapılmaktadır. Bu çalışma ile web sunucu erişim kayıtlarının web kullanım madenciliği ile analizi için "Log Analiz" isminde bir yazılım geliştirilmiştir. Hazırlanan yazılım web sitesine ait çeşitli istatistikî bilgileri çıkarmakta ve apriori algoritması ile birliktelik kurallarını bulmaktadır. Log Analiz ile Gaziosmanpaşa Üniversitesi kurumsal web sitesine ait 15 günlük sunucu erişim kayıtları incelenmiş ve çeşitli analiz sonuçları elde edilmiştir. Böylece, web sitesine ait erişim kayıtları kolayca analiz edilebilecek ve atıl durumdaki veriler bilgiye dönüştürülebilecektir.

Anahtar Kelimeler— Web madenciliği, apriori, birliktelik kuralları, web kullanım madenciliği, log analiz

LOG Analysis: Access Log Files Analysis Software and GOP University Application

Abstract— During visit the web site of internet users traces leaving behind are kept on the access log files of the server. This data were analyzed with web mining is transformed into knowledge. In this study, a software is developed named as "LOG Analysis" to user access logs of the web server were analyzed with web usage mining. The developed software variety of statistical information retrieval and association rules finds with Apriori algorithm. 15-days user access logs belonging to the web site of Gaziosmanpasa University are examined with "LOG Analysis" and various analysis results were obtained. Thus, access logs of the web site can be easily analyzed and idle data can converted into knowledge.

Keywords— Web mining, apriori, association rules, web usage mining, log analysis

1. GİRİŞ

Teknolojinin gelişmesi ve ucuzlaması ile birlikte internet hayatın bir parçası haline gelmiş ve insanlar bilgi ihtiyaçlarının birçoğunu internet üzerinden gerçekleştirir olmuştur. Kurum veya kuruluşlar interneti reklam, e-ticaret, iletişim, bilgi-belge paylaşımı, bankacılık işlemleri ve eğitim başta olmak üzere çeşitli alanlarda kullanmaktadır ve bu amaçlarla web sitelerine sahiptir.

Web siteleri gerek kullanıcı gerekse de sahibi için keşfedilmeyi bekleyen çok önemli bilgiler içermektedir. İnternet üzerinde bulunan veya web sitelerinin kullanımı sonucu oluşan verilerin anlamlandırılarak bilgiye dönüştürülmesi için web madenciliği kullanılmaktadır.

Web madenciliği, web sitelerini ziyaret eden kullanıcıların davranışlarını inceleyerek web sitelerinin güncellenmesi veya geliştirilmesi, müşterilerin ilgi alanları, reklam alma, pazarlama stratejileri oluşturma, sayfa kullanım dağılımlarını belirleme gibi birçok konuda karar verilmesini sağlayan bilgileri sunar.

Web kullanım madenciliği ile ilgili yapılmış araştırma projelerinde birbirinden farklı birçok yazılım geliştirilmiştir. Bu araştırma projelerini uygulama alanlarına, kullandıkları veri kaynaklarına ve veri tiplerine göre sınıflandırmak mümkündür. Tablo 1'de verilen ve geliştirilen yazılımların çoğu sunucu temelli verileri kullanmaktadır. Görüldüğü üzere yazılım projelerinin tümü kullanım verilerini, birkaçı ise kullanımın yanı sıra yapı, içerik veya profil verilerini kullanarak analiz yapabilmektedir. Tek kullanıcılı projeler genellikle kişiselleştirme uygulama alanını içermektedir. Çoklu site analizini destekleyen projelerde ise birden fazla web sitesinin kullanım verilerine kolayca erişebilmek için ya istemci ya da vekil sunucu seviyesinde giriş verileri kullanılmaktadır. Çoğu web kullanım madenciliği

projelerinde tek ve çok kullanıcılı siteler, web sunucu kayıtları gibi sunucu temelli kullanım verileri kullanılmaktadır.

Bu çalışmanın amacı, erişim kayıtlarını temizleyip veritabanına aktararak kolay kullanılabilir ve anlaşılır hale getirmek ve Gaziosmanpaşa Üniversitesi web sitesine ait erişim kayıtlarını inceleyerek sitenin analizini yapmaktır. Bu amaçla "Log Analiz" isminde .NET tabanlı yazılım geliştirilmiştir.

2. WEB MADENCİLİĞİ

Web madenciliği ilk olarak 1996 yılında Oren Etzioni tarafından ortaya atılmıştır. Bu bildiride Etzioni'ye göre [2] web madenciliği, veri madenciliği tekniklerini kullanarak www'de bulunan dosya ve servislerden otomatik olarak bilginin ayıklanması, ortaya çıkartılması ve analiz edilmesidir.

Tablo 1. Web kullanım madenciliği projeleri ve yazılımları [1]

Proje Adı	Uygulama Alanları	Veri Kaynağı			Veri Tipi				Kullanıcı		Site	
		Sunucu	Vekil	İstemci	Yapı	İçerik	Kullanım	Profil	Tek	Çok	Tek	Çok
WebSIFT(CTS99)	Genel	X			X	X	X			X	X	
SpeedTracer (WYB98,CPY96)	Genel	X					X			X	X	
WUM (SF98)	Genel	X			X		X			X	X	
Shahabi (SZAS97,ZASS97)	Genel			X	X		X			X	X	
Site Helper (NW97)	Kişiselleştirme	X					X		X		X	
Letizia (Lie95)	Kişiselleştirme			X			X		X			X
Web Watcher (JFM97)	Kişiselleştirme		X				X	X		X		X
Krishnapuram (NKJ99)	Kişiselleştirme	X					X			X	X	
Analog (YJGD96)	Kişiselleştirme	X					X			X	X	
Mobasher (MCS99)	Kişiselleştirme	X			X		X			X	X	
Tuzhilin(PT98)	İş	X					X			X	X	
SurfAid	İş	X					X			X	X	
Buchner (BM98)	İş	X					X	X		X	X	
WebTrend, Accrue	İş	X					X			X	X	
WebLogMiner (ZXH98)	İş	X					X			X	X	
WebLogMiner	Site Yenileme	X					X			X	X	
PageGather,SCML(PE98,PE99)	Karakterize etme	X			X		X			X	X	
Manley (Man97)	Karakterize etme	X					X			X		X
Arlitt (AW96)	Karakterize etme	X					X			X		X
Pitkow(PIT97,PIT98)	Karakterize etme	X		X			X			X		X
Almeida (ABC96)	Site Geliştirme	X					X			X		X
Rexford (CKR98)	Site Geliştirme	X	X				X			X	X	
Schechter (SKS98)	Site Geliştirme	X					X			X	X	
Aggarwal(AY97)	Site Geliştirme		X				X			X	X	

Web ortamındaki verilerin büyük olması kadar düzensiz olması da web madenciliğine ayrı bir önem kazandırmaktadır [3].

Web madenciliği ile işlenecek olan veri web sitesinin içerdiği bilgiler, sitenin yapısı, kullanıcıların ziyaretleri esnasında sunucu tarafından toplanan veriler ve ziyaretçilerin üyelik işleminde vermiş olduğu bilgilerden oluşmaktadır.

Web madenciliği çalışma alanlarının kapsamlı ve detaylı olması bu alanda düzenli bir sınıflandırmayı da gerektirmektedir. Web madenciliği ilk ortaya atıldığı dönemlerde Web İçerik Madenciliği ve Web Kullanım Madenciliği olmak üzere iki sınıfa ayrılmaktaydı. Web madenciliğinin yaygınlaşması ile birlikte Web Yapı Madenciliği de üçüncü bir sınıf olarak eklenmiştir [2,4].

Web içerik madenciliği, www'de bulunan içerik verisinden bilgi çıkarım işlemini gerçekleştirir [5]. Web yapı madenciliği, web sayfaları ve web siteleri arasındaki bağlantıları yani web yapı verisini inceleyerek bilgi çıkarım işlemini gerçekleştirir [5, 6]. Web log mining olarak da bilinen web kullanım madenciliği ise sunucu üzerinde tutulan ziyaret kayıt dosyalarından bilgi çıkarım işlemini gerçekleştirir.

2.1. Web Kullanım Madenciliği

Web sitesinin kullanım analizi için web kullanım verilerinden en yoğun ve en ilginç kullanıcı erişim örüntülerini keşfetmek ve anlamlı verileri çıkartmak için veri madenciliği tekniklerini uygulama sürecidir.

Web kullanım madenciliğinin ana veri kaynağını oluşturan web kullanım verisi web ve uygulama sunucusu üzerinde otomatik olarak toplanır. Genel olarak her bir erişim için erişim tarihi, saat, sunucu IP adresi, istemci IP adresi, istekte bulunulan web adresi, sayfa referansları, tarayıcı ve işletim sistemi bilgilerini içeren user-agent bilgisi tutulur. Tutulan parametre sayısı sunucu üzerinde yapılan konfigürasyon göre değişiklik gösterebilir.

Web kullanım madenciliği ön işlem, örüntü keşfi ve örüntü analizi olmak üzere 3 aşamada gerçekleştirilir [4]. Bu aşamalar Şekil 1'de gösterilmiştir.

2.1.1. Ön İşlem Süreci

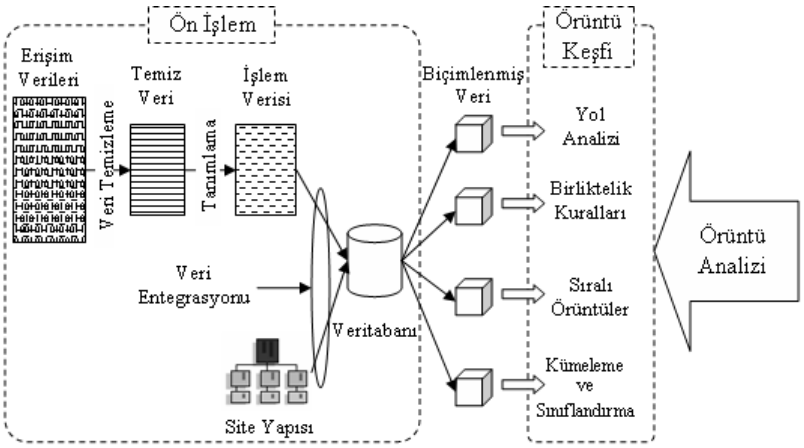
Sunucular üzerinde karmaşık ve düzensiz bir şekilde tutulan log dosyalarındaki verilerin analiz değeri olmayan ilişkisiz verilerden temizlenmesi, belirli bir biçime getirilmesi ve veritabanına aktarılması işlemi ön işlem sürecidir. Ön işlem süreci web kullanım madenciliğinin en önemli ve en uzun süren basamağıdır. Bu süreçte önemli olan verinin orijinalliğinin korunmasıdır.

Ön işlem süreci veri temizleme, kullanıcı tanımlama, oturum tanımlama, yol tamamlama ve biçimlendirme olmak üzere dört adımda gerçekleşir. Verilerin temizlenmesi, kullanıcı ve oturum tanımlama aşamalarında sezgisel (heuristic) teknikler kullanılmaktadır [8].

Veri Temizleme: Erişim kayıtları içerisindeki geçerli ve gerekli olan veriler alınmalı diğerleri temizlenmelidir [9]. Temizliğe ihtiyaç duyulan gereksiz veya alakasız üç tür veri vardır. Bunlar HTML dosya içerisine gömülü ek kaynaklar, robot istekleri ve başarısız isteklerdir.

Kullanıcı Tanımlama: Benzer kullanıcılara ait olan aktiviteleri belirlemek için kullanılır. Kimlik doğrulama veya kullanıcı taraflı çerezler olmaksızın kullanıcıları tanımlamak için IP adresi ile birlikte tarayıcı ve işletim sistemi bilgilerini tutan user-agent bilgisi de kullanılabilir.

Oturum Tanımlama: Bir oturum kullanıcının siteye girişi ile çıkışı arasındaki sürede gerçekleştirdiği aktiviteler grubu olarak tanımlanabilir. Bu nedenle oturum tanımlama işlemi, web oturumları içerisindeki her bir kullanıcının davranış ve aktivite kayıtlarının kümelenmesidir [8]. Kimlik doğrulama sistemi bulunmayan web sitelerinde oturum tanımlama işlemi için oturum süresi temelli, sayfada kalma süresi temelli ve referans temelli olmak üzere üç sezgisel yaklaşım bulunmaktadır [8, 10, 11].



Şekil 1. Web kullanım madenciliğinin uygulama adımları [7]

Yol Tamamlama: Erişim kayıtları vekil sunucuda tutuluyorsa veya site gezintisi esnasında ön bellekten sayfa ziyaretleri gerçekleşiyorsa log dosyaları içerisine kaydedilmeyen önemli erişimler vardır. Örneğin, site içerisinde gezinti yapan bir kullanıcı tarayıcı üzerinden geri düğmesi ile gezinti yaparsa veya vekil sunucudan istekte bulunulan sayfa sunucunun ön belleğinden gösterilirse bu erişimler log dosyası içerisinde yer almayacaktır. Yol tamamlamanın görevi erişim kayıtları içerisinde bulunan bu eksik referansları tamamlamaktır [12].

2.1.2. Örüntü Keşfi

Örüntü keşfi aşamasında ön işlem sürecinden sonra elde edilen düzenli ama anlamsız olan verilerden, veri madenciliği yöntemlerini kullanarak istenilen faydalı ve gerekli bilgilerin ortaya çıkarılması gerçekleştirilmektedir.

Web kullanım madenciliği örüntü keşfi aşaması için istatistiksel analiz, birliktelik kuralları, yol analizi, kümeleme, sınıflandırma ve sıralı örüntüler yaygın olarak kullanılmaktadır.

2.1.3. Örüntü Analizi

Örüntü analizi web kullanım madenciliğinin son adımıdır. Örüntü analizinin amacı bulunan örüntülerden ilginç olmayan kuralları, istatistikî bilgileri ya da örüntüleri elemektir [13, 7]. Genellikle örüntü analiz işlemi web madenciliği uygulamaları tarafından elde edilir. SQL, MySQL gibi veritabanı uygulamaları ve On-Line Analytical Processing (OLAP) yaygın olarak kullanılan bilgi sorgulama mekanizmalarıdır.

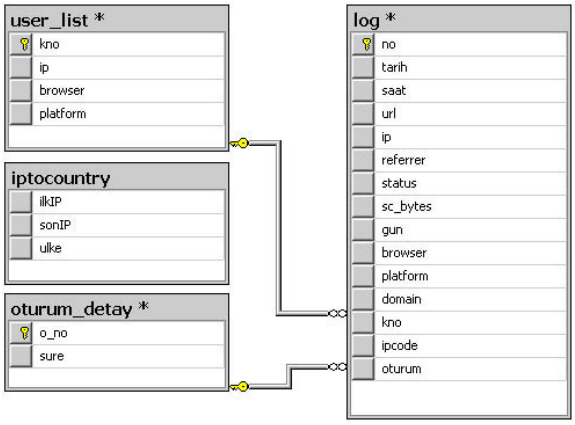
Örüntü analizi konusunda yapılmış birçok çalışma ve uygulamalar mevcuttur. Örneğin; Iocchi makale çalışmasında [14], web kullanım madenciliği uygulaması ile kullanımı kolay ve anlaşılır kullanıcı ara yüzü sayesinde kullanıcının istekleri ve seçimleri doğrultusunda örüntü analizi yapılabilmektedir. Web kayıt dosyalarının temizlenip, istatistikî bilgilerin elde edilmesini sağlayan eWebLog [15], NetIQ [16], Nihuo [17], Sarg [18] ve WebTrends [19] gibi birçok farklı program bulunmaktadır.

3. LOG ANALİZ PROGRAMI VE UYGULAMA

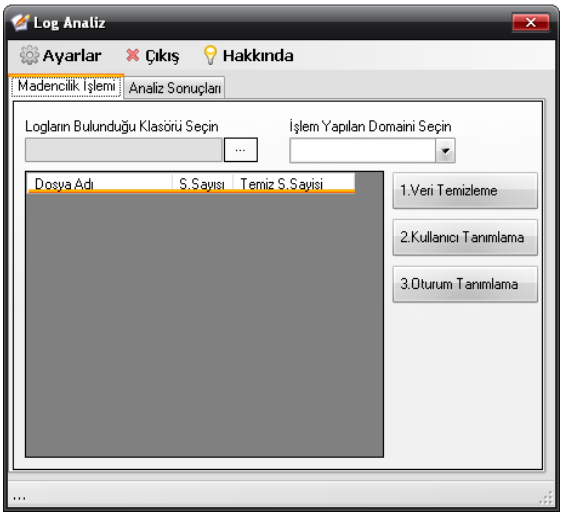
Log Analiz programı Visual C# .NET programlama dili ile hazırlanmış ve MSSQL Server 2005 Express veritabanı kullanılmıştır. Kullanılan veritabanının yapısı Şekil 2'de verilmiştir.

3.1. Yazılımın Özellikleri

Hazırlanan yazılım formlar yardımıyla tasarlanarak kullanımı kolaylaştırılmıştır. Kullanıcının analiz için özel bir kod bilmesine veya komut yazmasına gerek kalmamaktadır. Programa ait ekran görüntüsü Şekil 3'te verilmiştir.



Şekil 2. Log Analiz programının kullandığı veritabanı diyagramı



Şekil 3. Log Analiz programı açılış penceresi

Program Ayarlar, Çıkış, Hakkında menülerinden ve Madencilik İşlemi, Analiz Sonuçları sekmesinden oluşmaktadır.

3.1.1. Ayarlar Menüsü

Madencilik süresince ihtiyaç duyulan bilgilerin tanımlanması için kullanılır. Şekil 4'te ayarlar menüsünün içeriği gösterilmiştir.

Şekil 4. Ayarlar menüsü içeriği

SQL Server: Veritabanı sunucusuna bağlantı için gerekli ayarlamalar yapılır.

LOG Yapısı: Web kullanım madenciliği için kullanılan erişim kayıtları her zaman aynı formatta olmamaktadır. Bu seçenek ile analizi yapılacak erişim kayıtlarının formatı programa tanıtılmaktadır.

Dikkate Alınmayacak Sayfalar: Site tasarımından kaynaklı olarak erişim kayıtları içerisinde dikkate alınmayacak sayfaları tanımlamak için kullanılır.

Dikkate Alınacak Dosya Uzantıları: Bu seçenek ile erişim kayıtları içerisinde madencilik işlemine dahil edilecek dosya uzantıları belirlenmektedir.

Bot, Spider, Crawler Anahtar Kelimeler: Erişim kayıtları içerisinde madencilik sürecine dahil edilmeyecek olan robot isteklerin tanınması için kullanılacak anahtar kelime tanımlaması için kullanılır.

3.1.2. Madencilik İşlemi Sekmesi

Log analiz programı, web kullanım madenciliğinin ön işlem sürecini madencilik işlemi sekmesinde gerçekleştirmektedir.

Erişim kayıtları ve domain seçildikten sonra ön işlem sürecinin basamakları olan veri temizleme, kullanıcı tanımlama ve oturum tanımlama işlemlerini gerçekleştirir.

3.1.3. Veri Temizleme

Erişim kayıtlarının içerdiği verilerin tamamı madencilik süreci için gerekli veriler değildir. Bu nedenle, erişim kayıtları içerisindeki geçerli ve gerekli olan veriler alınmalı diğerleri temizlenmelidir.

Log analiz programının ayarlar kısmında tanımlanan dosya uzantıları dışında kalan dosyalara erişimler ve robot istekler için tanımlanan anahtar kelimeleri içeren satırlar gereksiz veri olarak belirlenmekte ve veritabanına eklenmemektedir.

Veritabanına aktarılırken geçerli olan bir veri satırının içerdiği verilerden date, time, cs-uri-stem, c-ip, cs-referer, sc-status, sc-bytes ve cs(User-Agent) verileri üzerinde aşağıda verilen biçimlendirmeler yapılmakta ve veritabanına aktarılmaktadır.

- Date verisi log tablosu üzerinde tarih sütununa ve tarihin gün karşılığı bulunarak gun sütununa eklenmektedir.
- Log tablosu üzerinde time verisi saat sütununa, cs-uri-stem verisi url sütununa, sc-status verisi status sütununa, sc-bytes verisi sc_bytes ve cs-referer verisi referrer sütununa eklenmektedir.
- cs(User-Agent) verisi üzerinden ziyaretçinin kullandığı tarayıcı ve işletim sistemi tespit edilerek log tablosu üzerine browser ve platform sütununa eklenmektedir.
- c-ip verisi üzerinden kullanıcının IP adresinin sayısal karşılığı bulunarak log tablosu üzerinde ipcode sütununa eklenmektedir. Dönüşüm için kullanılan kod parçası Şekil 5'de verilmiştir.

```
String ip_adresi="193.140.80.2";
String[] ip= ip_adresi.Split('.');
Double ipcode =16777216 * Convert.ToDouble(ip0]) +
65536 * Convert.ToDouble(ip[1]) +
256*Convert.ToDouble(ip[2]) +
Convert.ToDouble(ip[3]);
```

Şekil 5. IP adresini sayısal değere dönüştürmek için kullanılan kod parçası

3.1.4. Kullanıcı Tanımlama

Kullanıcı tanımlama işlemi için ziyaretçinin IP adresi, kullandığı işletim sistemi ve tarayıcı bilgileri kullanılmaktadır. Bu üç bilgisi aynı olan erişimler tek bir kullanıcı olarak tanımlanmaktadır.

Kullanıcı tanımlama programlama tarafında yapıldığında veritabanının sorgulaması ve oluşturulan kullanıcıların tekrar veritabanına aktarılması için veritabanına birçok kez bağlantı kurmak gerekir ve programın hızında ciddi yavaşlamalar meydana gelir. Bu sorunun önüne geçmek için veritabanı üzerinde *user_create* isminde yordam tanımlanarak kullanıcı oluşturma işlemi bu yordam yardımıyla yapılmıştır. Böylece veritabanına bir kez bağlantı yapılarak programda performans artışı sağlanmıştır. Oluşturulan *user_create* yordamına ait T-SQL ifadesi Şekil 6'da verilmiştir.

```
CREATE PROCEDURE user_create AS
INSERT INTO user_list SELECT ip,browser,platform
FROM log GROUP BY ip,browser,platform
UPDATE log SET kno = (SELECT top 1 kno FROM
user_list WHERE log.ip = ip AND log.browser =
browser AND log.platform = platform)
RETURN
```

Şekil 6. user_create yordamına ait T-SQL kodları

3.1.5. Oturum Tanımlama

Hazırlanan yazılımda oturum tanımlama için oturum süresi temelli yaklaşım kullanılmıştır. Bu yaklaşıma göre her bir oturum belirlenen eşik değerini(θ) aşmamalıdır. Catledge ve Pitkow [20] ziyaretçilerin siteden ayrılma sürelerini ölçmüştür ve bu değer 1.5 standart sapma ile 25.5 dakikadır. Bu eşik değeri birçok çalışmada ortalama 30 dakika olarak kullanılmaktadır. Herhangi bir kullanıcı için oturumun başlangıç zamanı t0 ve bitiş zamanı tn olarak düşünüldüğünde, $t_n - t_0 \leq \theta$ şartını sağlayan tüm erişim aktiviteleri aynı oturum içerisinde değerlendirilir [21].

Erişimler için oluşturulan oturumlar log tablosu içerisinde oturum sütununa, tanımlanan oturumların numarası ve oturumun süresi oturum_detay tablosuna eklenmiştir. Bir oturum içerisinde tek bir sayfa ziyareti gerçekleştirilmişse oturum süresi 0 olarak tanımlanmıştır.

3.1.6. Analiz Sonuçları Sekmesi

Analiz sonuçları sekmesinde ön işlem sürecinden sonra elde edilen ve veritabanına aktarılan verilerden çeşitli istatistiki bilgiler elde edilmekte ve veri madenciliği tekniklerinden apriori algoritması ile birlikte ziyaret edilme olasılığı yüksek sayfalar tespit edilmektedir.

Veritabanından elde edilen verilerin grafiksel olarak gösterimi için Visual studio 2005 ile birlikte Microsoft Chart Controls for Microsoft .NET Framework 3.5 kullanılmıştır [22].

Log analiz programı ile elde edilebilecek bilgiler Şekil 7'de analiz sonuçları sekmesinde gösterilmiştir.

Şekil 7. Log analiz programı analiz sonuçları sekmesi

Apriori algoritmasına ait sözdekod Şekil 8'de verilmiştir [23].

4. UYGULAMA SONUÇLARI

Hazırlanan Log analiz programı Gaziosmanpaşa Üniversitesi web sitesine ait 15 günlük erişim kayıtları üzerinde uygulanmıştır. Kullanılan verilere ve elde sonuçlara ait çeşitli bilgiler Tablo 2'de verilmiştir.

4.1. Genel Bakış

Genel bakış seçeneği ile madencilik süreci sonrası erişimlere ait elde edilen çeşitli bilgiler listelenmektedir. Giriş verileri sonucu elde edilen genel bakış sonuçları Şekil 9'da verilmiştir.

Giriş: D, veritabanı hareketleri; *min_sup,* minimum destek eşiği.
Çıkış: L, D'de yer alan sık geçen öğe kümeleri
Metod:
L_1 = find_frequent_1-itemsets(d);
for(k =2; $L_{k-1} \neq \emptyset$; k++) {
    C_k = aprirori_gen(L_{k-1});
    for each transaction $t \in D$ {
        C_t = subset(C_k, t);
        for each candidate $c \in C_t$
            c.count++; }
    $L_k = \{ c \in C_k \mid c.count \geq min\_sup \}$
} return $L = \cup_k L_k$;
Procedure apriori_gen(L_{k-1}:frequent(k – 1)-itemsets)
for each itemset $I_1 \in L_{k-1}$
  for each itemset $I_2 \in L_{k-1}$
    if ($I_1[1]=I_2[1]$)∧($I_1[2]=I_2[2]$)∧…∧($I_1[k-2]=I_2[k-$∧($I_1[k-1]=I_2[k-1]$) then {
      $c=I_1 \bowtie I_2$;
    if has_infrequent_subset(c, L_{k-1}) then
      delete c;
    else add c to C_k; }
return C_k;
Procedure has_infrequent_subset(c:candidate k-itemset; L_{k-1}: frequent(k -1)-itemsets);
  for each (k-1)-subset s of c
    if $s \notin L_{k-1}$ then return TRUE;
return FALSE;

Şekil 8. Apriori algoritması sözdekodu

Tablo 2. Kullanılan ve elde edilen verilere ait çeşitli bilgiler

Veriye ait özellik	**Değer**
Kullanılacak verinin tarih aralığı	01.11.2010-15.11.2010
Erişim dosyası sayısı	15 adet
Toplam veri boyutu	2.08 GB
Logların içerdiği toplam satır sayısı	8 531 350 satır
Ön işlem sonrası satır sayısı	172 808 satır
Veritabanı boyutu	88.5 MB

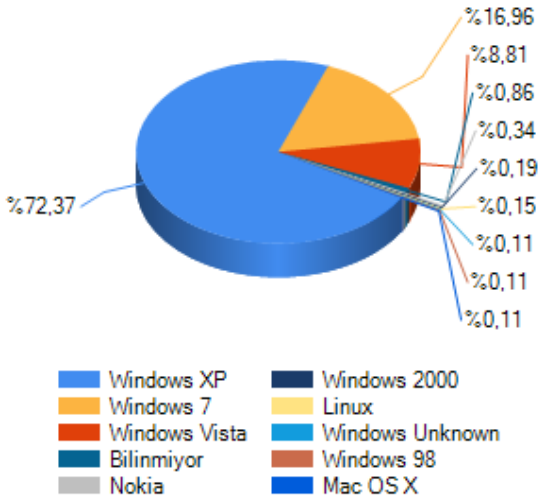
Genel Bakış

Analiz	Sonuç
İlk Erişim Zamanı	01.11.2010 00:00:24
Son Erişim Zamanı	15.11.2010 23:55:56
Toplam Ziyaret	172808
Günlük Ortalama Ziyaret	11520
Ziyaret Sayısı	78588
Farklı Kişi Ziyaret Sayısı	33365
Sitede Geçirilen Ortalama Süre(Saniye)	238
Hemen Çıkma Oranı	23,11%
Başarısız Erişimler	2323
Görüntülenen Sayfa Sayısı	361
Günlük Ortalama Görüntülenen Sayfa Sayısı	24
Hafta İçi Ziyaret Sayısı	148596
Hafta Sonu Ziyaret Sayısı	24212
Haftanın En Aktif Günü	Pazartesi
Haftanın En Pasif Günü	Pazar
Günün En Aktif Saati	09:00 - 09:59
Günün En Pasif Saati	03:00 - 03:59

Şekil 9. Genel bakış sonuçları

4.2. OS Dağılımı

Ziyaretçilerin kullanmış olduğu işletim sistemlerinin dağılımı Şekil 10'da verilmiştir.



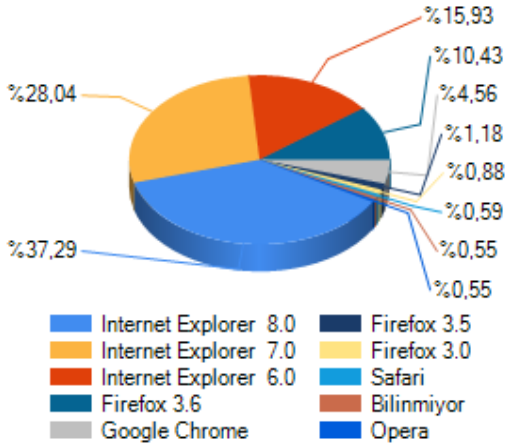
Şekil 10. İşletim sistemi dağılımları

4.3. Tarayıcı Dağılımı

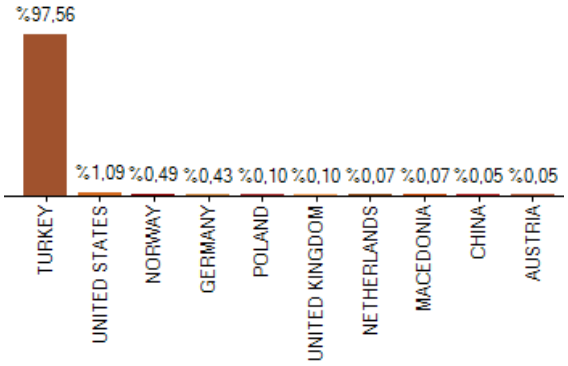
Ziyaretçilerin kullanmış olduğu internet tarayıcıların dağılımı Şekil 11'de verilmiştir.

4.4. Ülke Dağılımı

Web sitesini ziyaret eden kullanıcıların ülke dağılımı Şekil 12'de verilmiştir.



Şekil 11. Tarayıcı dağılımları

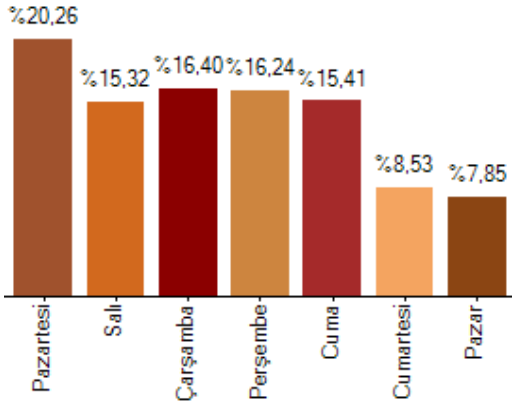


Şekil 12. Ülke dağılımı

Analiz sonucuna göre erişimlerin neredeyse tamamı kurumun bulunduğu ülkeden gerçekleştirilmektedir.

4.5. Günlük Dağılım

Haftanın günlerine göre ziyaret oranları Şekil 13'de verilmiştir.



Şekil 13. Günlük dağılımlar

Web sitesine erişimler en yoğun olarak haftanın ilk günüdür. En düşük erişimler hafta sonlarında

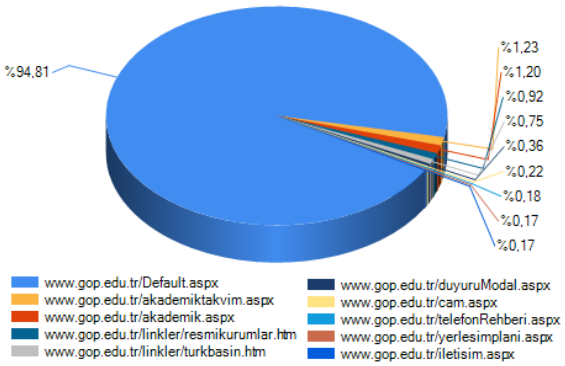
gerçekleştirilmiştir. Bu sonuç erişimlerin büyük çoğunluğunun kamu kurum ve kuruluşlarından yapıldığını göstermektedir. Haftanın ilk gününde yoğunluk daha fazla olduğu için erişimlerin önemli bir kısmının sitenin sahibi kurum üzerinden gerçekleştirildiği düşünülebilir.

4.6. Aylık Dağılım

Aylara göre ziyaret oranlarını bulmaktadır. Uygulamada sadece 15 günlük veri kullanıldığı için sonuçlar verilmemiştir.

4.7. En İyi Giriş Sayfaları

En iyi giriş sayfaları ziyaretçilerin site üzerinde gezintiye başladığı ilk sayfayı ifade etmektedir. Kullanılan verilere göre en iyi giriş sayfaları Şekil 14'te verilmiştir.

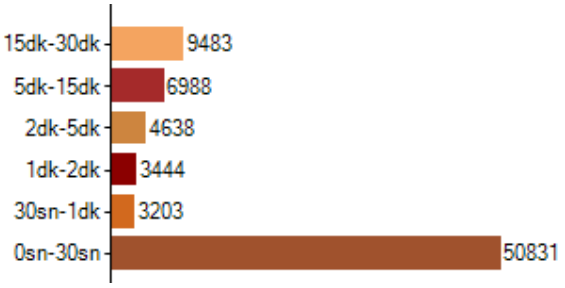


Şekil 14 En iyi giriş sayfaları

Sonuçlara göre, siteye erişimlerin neredeyse tamamı ana sayfa üzerinden gerçekleştirilmektedir. Bu durum siteye erişimlerin doğrudan veya arama motorları ile yapıldığını göstermektedir.

4.8. Ziyaret Süreleri

Şekil 15, web sitesini ziyaret eden kullanıcıların site üzerinde geçirdikleri sürelere göre dağılımını göstermektedir.

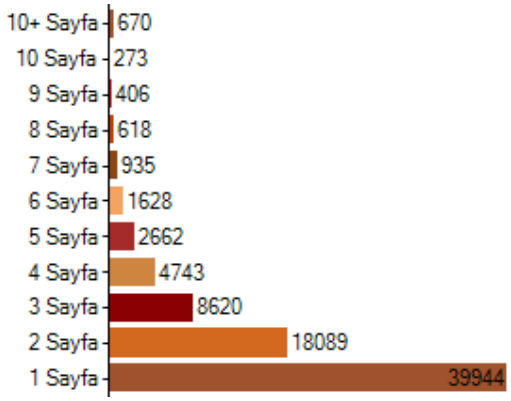


Şekil 15. Ziyaret süreleri dağılımı

Ziyaretlerin büyük bir kısmı 0-30 sn arasında sürmektedir. Bu da gösteriyor ki ziyaretçiler sitenin ana sayfasında bulunan duyuru ve haber gibi bilgileri inceledikten sonra siteden ayrılmaktadır. Bu durum kullanıcıların büyük çoğunluğunun kurum personeli veya öğrenci olduğunu göstermektedir.

4.9. Ziyaret Derinliği

Şekil 16, ziyaretçilerin her bir oturumda ziyaret ettiği sayfa sayısının dağılımını göstermektedir. 10 sayfadan fazla olan oturumlar tek bir sütun üzerinde gösterilmiştir.

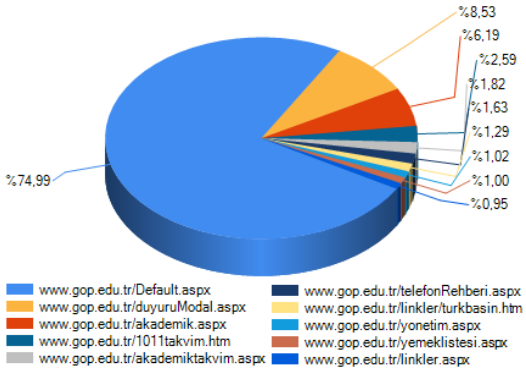


Şekil 16. Ziyaret derinliği dağılımı

Ziyaret süreleri ve en iyi giriş sayfaları analizleri ile elde edilen sonuçlar burada elde edilen sonuçlarla tekrar onaylanmıştır. Yani kuruma ait web sitesinin ana sayfası ziyaretlerin çok önemli bir kısmını oluşturmaktadır. Ana sayfa dışındaki sayfalara erişim çok düşüktür. Kurumun web sitesi üzerinde duyuru ve haber kısmı ana sayfada bulunmaktadır. Bu durum sitenin büyük çoğunlukla haber ve duyurular için ziyaret edildiğini ve ziyaretçilerin kurum personeli veya öğrencileri olduğunu göstermektedir.

4.10. İlk 10

Şekil 17, web sitesi içerisinde en yoğun olarak kullanılan ilk 10 sayfayı göstermektedir.



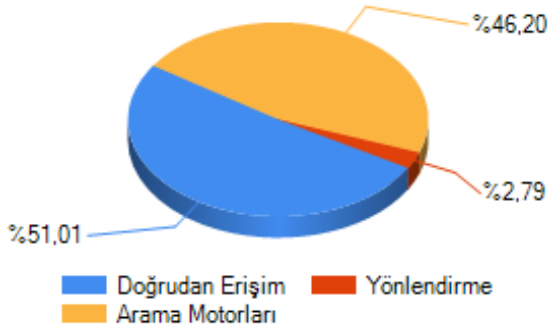
Şekil 17. İlk 10 dağılımı

Analiz sonuçlarına göre kuruma ait web sitesi büyük çoğunlukla duyuru, haber ve akademik takvim için ziyaret edilmektedir. Web sitesinin neredeyse hiç ziyaret edilmemiş sayfaları bulunmaktadır.

4.11. Trafik Dağılım

Şekil 18'deki trafik dağılımı, web sitesini ziyaret eden kullanıcıların siteye hangi kaynaktan ulaştığını grafiksel olarak göstermektedir. Trafik dağılımı doğrudan erişim, arama motorları ve yönlendirme olmak üzere 3 gruba ayrılmıştır.

Erişim kayıtları içerisinde bulunan cs (Referer) verisi ziyaret edilen sayfaya hangi adresten gelindiğini göstermektedir. Eğer bu bilgi boş ise siteye doğrudan erişildiğini göstermektedir. Yani kullanıcının tarayıcı adres çubuğuna adresi yazarak yaptığı erişimlerdir. Boş olan referans verisi haricindeki veriler arama motorlarının anahtar kelimelerine (google, bing, search v.b.) göre kontrol edilmekte bunlardan birisi var ise arama motorları grubu içerisine alınmaktadır. Bu iki durum dışındaki erişimler ise yönlendirme olarak düşünülmekte ve yönlendirme grubuna eklenmektedir.

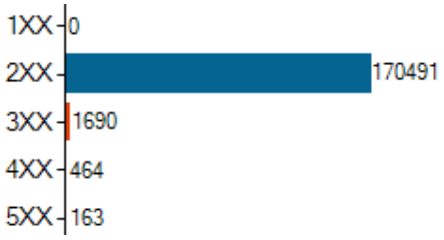


Şekil 18. Trafik dağılımı

Trafik dağılımı analizine göre doğrudan erişim yani kurumun web adresini adres çubuğu üzerine yazarak yapılan erişim sayısı ziyaretlerin yarısından daha fazlasını oluşturmaktadır. Web sitelerinde gezinti yapan kullanıcılar kendisi için öncelikli olmayan sitelerin adresini yazmaktansa arama motorları ile siteye erişim sağlamaktadır. Burada elde edilen doğrudan erişim oranının büyük çoğunluğu kurumla doğrudan veya dolaylı yoldan bağlantısı olan ziyaretçiler ile oluşturulmuştur. Arama motorları ile erişim önemli bir orana sahiptir. Bu oranı sitenin adresini bilmeyen kullanıcılar oluşturduğu ve aynı zamanda kurum içerisindeki kullanıcılarında bir kısmının arama motorlarını kullandığını göstermektedir.

4.12. Durum Kodu Dağılımı

Web sitesine yapılan erişimlerin tamamı başarılı olarak gerçekleşmez. Sunucudan veya kullanıcıdan kaynaklı hatalardan dolayı erişim gerçekleşmemiş olabilir. Bu tür durumlar da erişim kayıtları içerisinde kayıt altına alınmaktadır.



Şekil 19. Durum kodu dağılımı

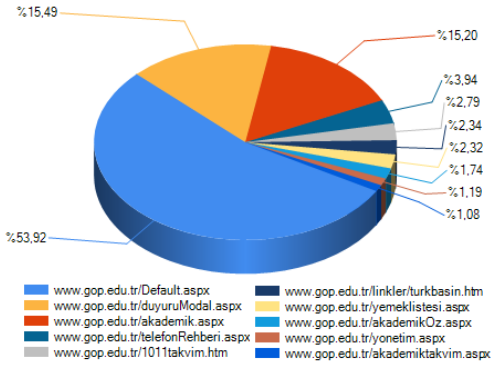
Elde edilen analiz sonucuna göre 5xx yani sunucu kaynaklı hataların değeri küçüktür. Bu da sunucu performansının iyi olduğunu göstermektedir.

4.13. Alt Domain Analizi

Uygulama verileri tek bir domain'e ait olduğu için sonuçlar verilmemiştir. Domain sayısı birden fazla olduğu durumlarda sonuçlar incelenebilir.

4.14. Çıkış Sayfaları

Şekil20, ziyaretçilerin siteden ayrıldıkları ilk 10 sayfayı göstermektedir.



Şekil 20. Çıkış sayfalarının dağılımı

Burada elde edilen sonuçların en iyi giriş sayfaları sonuçları ve ziyaret derinliği sonuçları ile bağlantısı bulunmaktadır. Yani ziyaretçiler büyük çoğunlukla sitenin ana sayfasını ziyaret etmiştir. Bu durum da ziyaretçilerin kurum personeli veya kurum öğrencisi olduğunu göstermektedir.

4.15. Apriori

Apriori algoritması birliktelik kuralı çıkarım algoritmaları içerisinde en fazla bilinen ve kullanılan algoritmadır. Sık geçen öğeleri bulmak için birçok kez veritabanını taramak gerekir. Bu taramalar aşamasında apriori algoritmasının birleştirme, budama işlemleri ve minimum destek ölçütü yardımı ile birliktelik ilişkisi olan öğeler bulunur. Apriori algoritması k+1 adet sık geçen öğe kümesini bulmak için k adet sık geçen öğe kümesine ihtiyaç duyar [23].

Sık geçen öğe kümelerini bulmak için veritabanını ilk taramada bir elemanlı minimum destek metriğini sağlayan sık geçen öğe kümeleri bulunur ve L_1 şeklinde gösterilir. L_1, L_2'yi ve L_2 ise L_3'ü bulmak için kullanılır. İzleyen taramalarda bir önceki taramada bulunan sık geçen öğe kümeleri aday kümeler (C_k) adı verilen yeni potansiyel sık geçen öğe kümelerini üretmek için kullanılır. Aday kümelerin destek değerleri tarama sırasında hesaplanır ve aday kümelerden minimum destek metriğini sağlayan kümeler o geçişte üretilen sık geçen öğe kümeleri olur. Sık geçen öğe kümeleri bir sonraki geçiş için aday küme olurlar. Bu süreç yeni bir sık geçen öğe kümesi bulunmayana kadar devam eder.

Genellikle alışveriş uygulamalarında kullanıldığı için market sepet analizi olarak da bilinmektedir. Bu yöntemdeki amaç bir küme içerisindeki nesnelerin birbirleri ile olan bağlarının tespit edilmesidir.

Sepet analizinde mallar arasındaki bağıntı, destek ve güven değerleri aracılığıyla hesaplanır. Destek veri içerisinde bu bağıntının ne kadar sık olduğunu, güven de X ürününü almış bir kişinin hangi olasılıkla Y ürününü alacağını ifade eder. Bağıntının önemli olabilmesi için her iki değerin de olabildiğince büyük olması gerekir.

X ve Y farklı ürünler olmak üzere, X ürünü için destek tüm alışverişler içinde X ürününün oranıdır. |X|, X ürünü içeren alışverişlerin sayısını, |D| yapılan tüm alışverişlerin sayısını göstermek üzere;

$$\text{Destek}(X) = \frac{|X|}{|D|} \quad (1)$$

X ve Y ürünleri için destek, X ve Y ürünlerini içeren alışveriş sayısı olmak üzere;

$$\text{Destek}(X \Rightarrow Y) = \frac{|X.Y|}{|D|} \quad (2)$$

X ve Y ürünleri için güven ise;

$$\text{Güven}(X \Rightarrow Y) = \frac{\text{destek}(X.Y)}{\text{destek}(X)} \quad (3)$$

Örneğin, bir X ürünü satın alan müşteriler aynı zamanda Y ürünün de satın alıyorlarsa, bu durumun birliktelik kuralı ile gösterimi;

$$X \Rightarrow Y \text{ [destek=\%2, güven=\%60]} \quad (4)$$

Buradaki destek ve güven ifadeleri, kuralın ilginçlik ölçüleridir. Sırasıyla, keşfedilen kuralın kullanışlığını ve doğruluğunu gösterir.

Log analiz programı apriori seçeneği Gaziosmanpaşa Üniversitesi web sitesine ait sayfalardan birlikte kullanılanları apriori algoritması ile tespit etmek için kullanılmıştır.

Apriori algoritmasının kullanacağı destek ve güven değeri program üzerinden apriori seçeneği seçildiğinde açılan pencere üzerinde Support(%) ve Confidence(%) kısmından belirlenir.

Algoritma içerisinde destek değerinin büyük olması apriori adımlarını azaltır ama elde edilen sonuç sayısı da azalır. Bu durumda elde edilen sonuç kümesi kullanışlı olmayacaktır. Önemli olan güven değerinin büyük olmasıdır. Çünkü güven değeri elde edilen kuralın doğruluğunu belirtmektedir.

Log analiz programında farklı destek değerleriyle elde edilen kurallar Tablo 3'te verilmiştir.

Tablo 3. Farklı destek değerleri ile elde edilen kurallar

Destek Değeri	Kural Sayısı	Sonuç
%25	0	Herhangi bir kural bulunamamıştır.
%20	1	Bir kural bulunmuştur ve elde kural anlamsızdır.
%10	2	Bulunan her iki kural da birbirini kapsamaktadır.
%5	5	Bulunan beş kural da tekli ilişkiler içermektedir.
%2	18	Bulunan kurallar tekli ve ikili kurallar içermektedir.
%1	12	Bulunan kurallar tekli, ikili ve üçlü ilişkiler içermektedir ama bazı ilişkiler kaybolmuştur.
%0.5	24	Bulunan kurallar tekli, ikili ve üçlü ilişkiler içermektedir ve bu ilişkilerin bazıları birbirini kapsamaktadır. Ayrıca gereksiz bazı ilişkiler bulunmuştur ve bu ilişkiler web sayfası için bir anlam ifade etmemektedir.

Farklı destek değerleri ile elde kurallara göre en sağlıklı sonuç %2 destek değeri için elde edilmiştir. %2 destek ve %75 güven ile elde edilen birliktelikler Şekil 21'de verilmiştir.

5. SONUÇ

Sunucu üzerinde tutulan erişim kayıtları herhangi bir metin editörü ile açılarak incelendiğinde herhangi bir anlam ifade etmeyen, karmaşık ve düzensiz bir yapıda olduğu görülecektir. Bu veriler web kullanım madenciliği ile analiz edilerek anlamlandırılmaktadır.

Bu çalışma da, web kullanım madenciliğinin tüm süreçlerini içeren ve ilgili erişim kayıtlarından çeşitli istatistikî bilgiler çıkartan bir yazılım tasarlanmıştır.

Apriori Rules

Consequent	Antecedent	Support (%)	Confidence (%)
www.gop.edu.tr/1011takvim.htm	www.gop.edu.tr/akademiktakvim.aspx	6,394	97,896
www.gop.edu.tr/akademiktakvim.aspx	www.gop.edu.tr/1011takvim.htm	6,309	99,221
www.gop.edu.tr/1011takvim.htm	www.gop.edu.tr/akademik.aspx www.gop.edu.tr/akademiktakvim.aspx	2,039	98,604
www.gop.edu.tr/akademiktakvim.aspx	www.gop.edu.tr/1011takvim.htm www.gop.edu.tr/akademik.aspx	2,031	98,981
www.gop.edu.tr/1011takvim.htm	www.gop.edu.tr/akademiktakvim.aspx	6,394	97,896
www.gop.edu.tr/akademiktakvim.aspx	www.gop.edu.tr/1011takvim.htm	6,309	99,221
www.gop.edu.tr/1011takvim.htm	www.gop.edu.tr/default.aspx www.gop.edu.tr/akademiktakvim.aspx	4,49	97,522
www.gop.edu.tr/akademiktakvim.aspx	www.gop.edu.tr/1011takvim.htm www.gop.edu.tr/default.aspx	4,417	99,121
www.gop.edu.tr/default.aspx	www.gop.edu.tr/akademik.aspx	20,461	96,560
www.gop.edu.tr/default.aspx	www.gop.edu.tr/duyurumodal.aspx	23,119	99,071
www.gop.edu.tr/default.aspx	www.gop.edu.tr/duyurumodal.aspx www.gop.edu.tr/akademik.aspx	2,883	99,641
www.gop.edu.tr/default.aspx	www.gop.edu.tr/linkler/turkbasin.htm	2,976	88,087
www.gop.edu.tr/default.aspx	www.gop.edu.tr/linkler.aspx	3,16	96,642
www.gop.edu.tr/linkler.aspx	www.gop.edu.tr/linkler/turkbasin.htm	2,976	84,783
www.gop.edu.tr/linkler/turkbasin.htm	www.gop.edu.tr/linkler.aspx	3,16	79,853
www.gop.edu.tr/linkler.aspx	www.gop.edu.tr/linkler/turkbasin.htm www.gop.edu.tr/default.aspx	2,621	92,300
www.gop.edu.tr/linkler/turkbasin.htm	www.gop.edu.tr/linkler.aspx www.gop.edu.tr/default.aspx	3,054	79,237
www.gop.edu.tr/default.aspx	www.gop.edu.tr/linkler.aspx www.gop.edu.tr/linkler/turkbasin.htm	2,523	95,897

Şekil 21. %2 destek ve %75 güven ile elde edilen birliktelikler

Hazırlanan yazılım ile

- Erişim kayıtlarına ait genel bilgiler,
- Ziyaretçilerin kullanmış olduğu işletim sistemleri dağılımı,
- Ziyaretçilerin kullanmış olduğu tarayıcı dağılımı,
- Ziyaret gerçekleştirilen ülkelerin dağılımı,
- Aylık erişim dağılımı,
- En iyi giriş sayfaları,
- Ziyaret süreleri,
- Ziyaret derinliği,
- En çok ziyaret edilen ilk 10 sayfa,
- Trafik dağılımı,
- Durum kodu dağılımı,
- Alt domain analizi,
- Çıkış sayfaları,
- Birlikte ziyaret edilen sayfalar gibi çeşitli sonuçlar elde edilmiştir.

Hazırlanan yazılım ile Gaziosmanpaşa Üniversitesi kurumsal web sitesine ait 15 günlük erişim kayıtları analiz edilmiş elde edilen sonuçlar uygulama sonuçları bölümünde verilmiştir. Hazırlanan yazılım en fazla 30.5 GB boyutunda 6 aylık erişim kayıtları üzerinde test edilmiş ve elde ettiği sonuçların geçerliliği çeşitli programlarla test edilmiştir.

Hazırlanan yazılım ile ön işlem sürecinden geçirilen veriler SQL veritabanına aktarılarak sonraki süreçlerin daha hızlı bir şekilde gerçekleştirilmesi sağlanmıştır.

Bu ve benzeri çalışmalar ile web site yöneticilerine web sitesinin geliştirilmesi veya yeniden tasarlanması için önemli bilgiler sunulmaktadır. Yapılan analizler sonucunda;

- Web sitesinin yoğun olarak ana sayfasının kullanıldığı iç kısımlarda kalan sayfaların çoğunlukla kullanılmadığı tespit edilmiştir.
- Ziyaret edilen sayfa sayıları incelendiğinde çoğunlukla duyuru, haber ve akademik takvimle ilgili sayfaların ziyaret edildiği görülmektedir ve bu durum ziyaretçilerin büyük çoğunluğunun mevcut öğrenciler ve personel olduğunu göstermektedir.
- Çeşitli destek ve güven değerleri ile yapılan birliktelik analizi sonuçlarının tamamında ana sayfa, akademik takvim ve aktif eğitim öğretim yılına ait akademik takvim birliktelikleri çıkmaktadır. Destek değeri düşürüldükçe uygulama sonuçları bölümünde verildiği gibi çeşitli birliktelikler çıkmaktadır. Ziyaretçiler aktif akademik takvime ana sayfa üzerinde bulunan akademik takvim linki ile ulaşmaktadır. Aktif döneme ait akademik takvim ana sayfaya yerleştirilerek erişim zamanı kısaltılabilir.

- İsteklerin oluşturduğu http durum kodu incelendiğinde sunucudan kaynaklı hataların oluşturduğu 5XX değerinin düşük olması genel olarak sunucunun sorunsuz çalıştığını göstermektedir.

KAYNAKLAR

[1] R. Cooley, **Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data**, Doktora Tezi, University of Minnesota, USA, 2000.

[2] O. Etzioni, "The World Wide Web: Quagmire or gold mine", *Communications of the ACM*, 39(11), 65-68, 1996.

[3] F. Gürcan, C. Köse, "Web İçerik Madenciliği ve Konu Sınıflandırması", **VI. İstatistik Günleri Sempozyumu**, Samsun, 1-5, 2008.

[4] R. Kosala, H. Blockeel, "Web mining research: a survey", *SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining ACM*, 2(1), 1–15, 2000.

[5] M. Kantardzic, **Data Mining:Concepts, Models, Methods and Algorithms**, John Wiley&Sons, New York, 2003.

[6] E. Belen, Ç. Özgür, B. Özakar, "WALA: Web Erişim Kütük Araştırmacısı", **9.Türkiye'de İnternet Konferansı**, İstanbul, 1-7, 2008.

[7] R. Cooley, B. Mobasher, J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", **In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)**, USA, 558–567, 1997.

[8] R. Cooley, B. Mobasher and J. Srivastava, "Data Preparation for mining World Wide Web Browsing Patterns", *Knowledge and Information Systems*, 1(1), 1–27, 1999.

[9] H. Liu, V. Keselj, "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests", *Data & Knowledge Engineering*, 61(2), 304-330, 2007.

[10] M. Spiliopoulou, B. Mobasher, B. Berendt, M. Nakagawa, "A framework for the evaluation of session reconstruction heuristics in web-usage analysis", *INFORMS Journal on Computing*, 15(2), 171–190, 2003.

[11] B. Berendt, B. Mobasher, M. Spiliopoulou, J. Wiltshire, "Measuring the accuracy of sessionizers for web usage analysis", **Proceedings of the Workshop on Web Mining at the First SIAM International Conference on Data Mining**, Chicago, 1-8, 2001.

[12] L. Chaofeng, "Research and Development of Data Preprocessing in Web Usage Mining", **International Conference on Management Science and Engineering**, China, 1311-1315, 2006.

[13] J. Srivastava, R. Cooley, M. Deshpande, P. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *SIGKDD Explorations*, 1(2), 12-23, 2000.

[14] L. Iocchi, "The Web OEM approach to Web Information Extraction", *Journal of Network and Computer Applications*, 22(1), 259-269, 1999.

[15] İnternet: eWebLog Analyzer, http://www.esoftys.com, 2010.

[16] İnternet: NetIQ Ssytem Management, http://www.netiq.com, 2010.

[17] İnternet: Web Log Analyzer-Nihuo, http://www.nihuo.com, 2010.

[18] İnternet: SARG, http://sarg.sourceforge.net, 20.10.2010.

[19] İnternet: WebTrends Teh Global Leader in Mobile and Social Analytics, http://www.webtrends.com, 2010.

[20] L. Catledge, J. Pitkow, "Characterizing browsing behaviors on the world wide web", *Computer Networks and ISDN Systems*, 27(6), 1065-1073, 1995.

[21] B. Liu,"Web Usage Mining", **Web Data Mining: Exploring Hyperlinks, Contents and Usage Data**, Springer Press, New York, 2006.

[22] İnternet: Microsoft Chart Controls for Microsoft .NET Framework 3.5, http://www.microsoft.com, 2010.

[23] J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Morgan Kaufmann Publishers, San Francisco, 2001.